1. Install Python + pip (if missing)

```
sudo yum update -y
sudo yum install python3 python3-pip -y
```

Verify:

```
python3 --version
pip3 –version
```

2. Install Required Packages

```
pip3 install --upgrade kaggle kagglehub awscli
export PATH=$PATH:~/.local/bin
```

Verify:

```
kaggle --version
aws –version
```

3. Kaggle API Setup

```
ls                              - kaggle.json
```

4. Move to correct folder

```
mkdir -p ~/.kaggle
mv kaggle.json ~/.kaggle/
chmod 600 ~/.kaggle/kaggle.json
kaggle datasets list
```

5. Check S3 Access (IAM role):

```
aws s3 ls
```

6. Clean Old Broken Downloads:

```
rm -rf ~/.cache/kagglehub
rm -rf flight_tab_only

rm -f *.zip
```

7. Create Python Pipeline Script:

```
cat > flight_tab_to_raw.py << 'EOF'
import kagglehub
import shutil
import subprocess
from pathlib import Path

DATASET = "flnny123/mfddmulti-modal-flight-delay-dataset"
WORK = Path("flight_tab_only")
S3_PATH = "s3://airport-airline-operations-analytics-platform/raw/"

print("STEP 1 — Downloading dataset...")
base_path = Path(kagglehub.dataset_download(DATASET))
print("Dataset extracted at:", base_path)

src = base_path / "Aeolus" / "Flight_Tab"
if not src.exists():
    raise RuntimeError("Flight_Tab folder not found")

print("STEP 2 — Copying Flight_Tab...")
shutil.copytree(src, WORK, dirs_exist_ok=True)

print("STEP 3 — Uploading to S3 raw folder...")
subprocess.check_call([
    "aws","s3","cp",
    str(WORK),
    S3_PATH,
    "--recursive"
])

print("STEP 4 — Count uploaded files:")
subprocess.call(
    f"aws s3 ls {S3_PATH} --recursive | wc -l",
    shell=True
)

print("DONE — RAW ingestion complete")
EOF
```

8. Run Pipeline

```
python3 flight_tab_to_raw.py
```

```
aws s3 ls s3://airport-airline-operations-analytics-platform/raw/flight_tab/ --recursive –summarize
```

9. Clean EC2 Disk (After Success)

```
rm -rf ~/.cache/kagglehub
rm -rf flight_tab_only
```